ABSTRACT OF THE DISCLOSURE

[0054] One embodiment of a distributed memory module cache includes tag memory and associated logic implemented at the memory controller end of a memory channel. The memory controller is coupled to at least one memory module by way of a point-to-point interface. The data cache and associated logic are located in one or more buffer components on each of the memory modules. The tag look-ups are performed in parallel with the memory module decodes. This improves latency for cache hits without penalizing the latency for cache misses.